AUTOMATED TRANSFER OF INFORMATION FROM PAPER DOCUMENTS TO COMPUTER-ACCESSIBLE MEDIA

by

Amar Gupta
Maher Kallel

Working Paper #2113-89-MS

AUTOMATED TRANSFER OF INFORMATION FROM PAPER DOCUMENTS TO COMPUTER-ACCESSIBLE MEDIA

by

Amar Gupta
Maher Kallel

Working Paper #2113-89-MS

PREFACE

In most organizations, the bulk of the information is maintained on paper. Paper based documents are unfortunately difficult to maintain and to update, and retrieving information from paper-based documents is a labor-intensive endeavor.

Transfer of information from paper documents to computer-accessible media, by methods other than manual keying, is indeed an interesting proposition. The technology in this area has improved significantly in recent years, and there are many applications which can benefit from such advance in the immediate future.

As is characteristic of most evolving fields, there are still gaps in what is expected of "reading machines" and what they are able to do today. Literature published by the vendors in this area contains specifications for their products, but these specifications are not easy to analyze on a comparative basis for specific user applications. Also, there is no benchmark test that are commonly accepted in this field.

In order to mitigate some of the weaknesses mentioned above, researchers at MIT's Sloan School of Management have launched an initiative to evaluate all major approaches and significant products with a view to determining overall trends. These researchers have also created a scheme for classifying paper documents based on their quality and their content. Using their classification scheme, it is possible to look at a document and quickly determine in which year off-the-shelf products either became available or will become available to "read" that document.

This paper is still being revised and enlarged. Comments and suggestions are solicited, and should be addressed to Dr. Amar Gupta, Principal Research Associate, Room E40-265, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139 (Telephone # [617] 253-8906).

AUTOMATED TRANSFER OF INFORMATION FROM PAPER DOCUMENTS TO COMPUTER-ACCESSIBLE MEDIA

Amar Gupta and Maher Kallel
Massachusetts Institute of Technology
Cambridge, MA 02139

1. INTRODUCTION

The concept of automated transfer of information from paper documents to computer-accessible media dates back to 1954 when the first Optical Character Recognition (OCR) device was introduced by Intelligent Machines Research Corporation. By 1970, approximately 1000 readers were in use and the volume of sales has grown to one hundred million dollars per annum. In spite of these early developments, scanning technology has so far been utilized in highly specialized applications only.

The lack of popularity of automated reading systems stems from the fact that commercially available systems have been generally unable to handle documents as prepared for human use. The constraints placed by such systems have served as barriers severely limiting their applicability. In 1982, Ullmann [1] observed:

"A more plausible view is that in the area of character recognition some vital computational principles have not yet been discovered or at least have not been fully mastered. If this view is correct, then research into the basic principles is still

needed in this area."

Research in the area of automated reading is being conducted as part of the wider field of pattern recognition. Less than one-tenth of the papers published in this latter field deal with character recognition. This research is slowly leading to new general purpose products. The overall capabilities of systems which are either available today, or are likely to become available in the next ten years, are analyzed in this paper.

2. CLASSIFICATION OF MACHINES

"Reading Machines" are usually classified into four major categories as follows:

2.1 Document Readers

This category of readers was the first to be introduced. Developed during the sixties and seventies, these machines are oriented towards transaction-processing applications such as billing and form processing. The source document is prepared in a rigid format using a stylized type font and the character set is limited to numeric characters plus a few special symbols and sometimes alphabetic letters. Document readers usually have a wide degree of tolerance in order to cope with material of poor quality produced by high speed printers. The speed of

processing is very high, typically between 400 and 4000 characters per second. The price, too, is high, usually several hundred thousand dollars.

2.2 "Process Automation" Readers

The main goal of these readers is to control a particular physical process. A typical application of such a reader machine is automatic sorting of postal letters. Since the objective in this case is to direct the piece of mail into the correct sorting bin, whether or not the recognition process results in the correct decision for every single character is not critical in this case. These types of readers are designed for specific applications in view, and are not considered in this paper.

2.3 Page Readers

These reading machines were originally developed to handle documents containing normal typewriter fonts and sizes. Initially intended for use in the newspaper industry and the publishing industry, these machines were designed with the capability to read all alphanumeric characters. Until the late seventies, devices of this type were quite restrictive in terms of their requirements for large margins, constrained

spacing, and very high quality of printing. Further, they accepted only a small number of specially designed fonts such as OCR-A and OCR-B. The reading speed for a monofont page was several hundred characters per second and the price of the machine itself was around one hundred thousand dollars per unit. The above situation was significantly altered by the introduction of the Kurzweil Omnifont Reader in 1978. As its name implies, this machine was able to read virtually any font. Since then, several other machines belonging to this category have been introduced. However, in spite of the availability of fairly powerful products, the use of page readers has continued to be restricted to specialized applications. The position is now beginning to change, however.

2.4 Hybrid Readers

The conventional distinction between page and document readers is gradually disappearing with the advent of hybrid readers, such as the Palantir Compound Document Processor which is able to process pages as well as documents. With this hybrid reader, the operator can specify the particular zones of the page to be read and the format of each zone. The higher flexibility, as compared to document processors, comes at the cost of speed which is only about 100 characters per second for

hybrid readers. The dividing line between the first three categories of readers will disappear over time.

The new generation of hybrid readers or general purpose readers offers a broad set of capabilities. These capabilities are analyzed in this paper. Machines that are available today are considered in the next section. These machines serve as the starting point for making predictions about the future.

3. CURRENT SET OF PRODUCTS

The products that have been examined in this paper include the following:

(a) Palantir CDP
(b) IOC Reader
(c) Kurzweil K4000
(d) Kurzweil Discover 7320
(e) CompuScan PCS 240
(f) Datacopy 730
(g) Microtek 300 C
(h) Abaton Scan 300/SF

Of the eight systems listed, the products from Palantir and Kurzweil (both Kurzweil K4000 and Kurzweil Discover 7320) can process all types of text, both typewritten and typeset. The remaining systems can read a limited number of text fonts.

Most low cost OCR systems available today, such as Datacopy

730, Microtek 300 C, and Abaton Scan 300/SF were originally developed to serve as image scanners for document processing. With recent developments in the area of desk-top publishing, these systems have been incorporated with an additional component of OCR software. Most of these systems can scan a page or a portion of a page as an image or as characters of text depending on the intent of the user. In some cases, the user can specify which zones within the input page contain text and which zones include graphic images. Some of the systems use separate phases for scanning text and images respectively. Most of these low cost scanners employ simple character recognition techniques such as matrix matching (described later in this paper). The basic hardware in each case is a desk-top scanner which depends on a host computer system for text processing. The host for these OCR systems is usually an IBM Personal Computer or compatible system.

At the other end, expensive systems such as Palantir CDP and Kurzweil products have well-developed hardware and software devoted to the tasks of text recognition and image scanning. Let us consider the Palantir CDP first. This system uses five dedicated Motorola 68000 processors and software written in C language to identify characters. Text processing is performed using a SUN work station or an IBM PC/AT compatible system. The first step is to read the input pages as images. The next step is to examine each image for presence of text. This process uses parallel processing technology and Artificial Intelligence-based

techniques to recognize textual material quickly and accurately. However, the system cannot automatically select separate areas for picking up text and images/pictures respectively. User intervention is needed for this process, as well as in the third step in which errors can be corrected interactively with the aid of system-supplied routines and bit-mapped image of unrecognized text areas.

As compared to the Palantir CDP, the Kurzweil K4000 system requires a highly skilled operator. The Kurzweil system is a dedicated one based on a minicomputer. It uses relatively old technology with a moving camera scanning lines over a page to identify letter within the text. Error correction and system training for unrecognizable fonts are performed simultaneously as the page is scanned. The training procedure limits the speed of the system. The Kurzweil Discover 7320 is a more recent product. It is a desk top unit for scanning pages and it can read all typewritten and type-set material. It uses an IBM PC/XT compatible system as the host system.

In order to analyze the overall trends, several products form the document processing and Computer Aided Design (CAD) areas deserve mention here. The document scanning systems (such as the one from Scan Optics) process large volumes of turn-around documents such as phone bills and checks. Significantly different from page readers, these systems offer several useful

(1) High speed image capturing and document processing capabilities using complex image capturing and analysis techniques and dedicated processors.

(2) Display of bitmaps of areas not recognized, allowing unrecognized and erroneous inputs to be corrected on an off-line basis.

(3) Recognition of hand-written text letters (within boxes, with no continuous flow of text).

The only general-purpose reader with some of the above capabilities is the Palantir CDP system which, though slower than document processors, captures the bit mapped image of the page being scanned, flags errors as described above, and interprets hand-written text (individual letters), though to a very limited extent.

In contrast to the emphasis on text in the systems described above, a Computer-Aided Design oriented scanning system captures a drawing in the form of its mapped image and then processes that image to generate an equivalent vector graphics file. This aspect is discussed later in Section 6.2 on Coalescence of Text, Image and Graphics Processing). It is expected that the techniques currently used in specialized systems will eventually become available in general-purpose systems.

We now turn our attention to the set of criteria used to specify the performance of scanning systems.

4. PERFORMANCE EVALUATION

Performance evaluation of a reading machine consists of several aspects. The conventional criteria for evaluation of a reader were, its accuracy and its speed. However, given the rapid evolution in the technologies used in the new generation of scanners, and their broad functionality, it becomes necessary to use a larger repertoire of evaluation criteria, as described in the following subsections.

4.1 Input Characteristics

The performance of a scanning system is heavily dependent on the characteristics of the input. For example, any reader will process a multifont, multisize document at a lower speed than a monofont, monosize document. The presence of graphics and the formatting of the document also affects the reading speed. Further, the quality of the print is a major factor affecting the accuracy of the system; broken and touching characters, low contrast, and skewed text result in high error rates and reject rates as well as in a significant reduction in the speed of reading. The speed and the error rate presented in the technical

documentation supplied by the vendors consider the characteristics in only one case - usually the perfect one.

4.2 Error and Reject Rates

The definition of accuracy is itself ambiguous. A character may be incorrectly recognized (which is an error) or is may be flagged as being unrecognizable (which is termed as a reject). There is a trade-off between errors and rejects. In fact, the error rate can be made arbitrarily small by increasing the rate of rejects. A reader may possess a very low error rate but it may flag or reject every character that does not offer a high probability of correctness as defined by the decision algorithm. As another example, consider a reader which recognizes the characters in "F16" correctly, but then flags them as this word is not in its built-in dictionary. As such, there is significant ambiguity in the definitions of errors and rejects. Also, the inability of readers to distinguish between ambiguous characters such as I (capital ai), l (el) and 1 may or may not be considered to be an error. The error rate of the entire recognition process is dependent not only on the functioning of the single character recognition subsystem but also on the level of preprocessing. For example, different kinds of

segmentation errors may occur and lines of text may be missed or misaligned in the case of a document containing several columns. Situations involving such missed or misaligned lines can be minimized by preprocessing.

Shurman [2] considers that with respect to performance evaluation, character recognition is a branch of empirical statistics. There is no reliable way of modelling the accuracy of a reading machine except by comparison with a standard set of norms. The impracticability of statistical modelling is due to the fact that the pattern generating process and its multivariate statistics are influenced by a number of barely controllable, application-dependent parameters.

4.3 Speed

In order to assess the capabilities of a system, one must consider not only the scanning speed but also the time spent in editing and correcting the document and the time spent in training the operators and the system itself where applicable. In the case of a document containing several graphs and multiple columns, the time spent in editing the document can be significantly greater than the scanning time. Consequently, it is difficult to obtain an accurate estimate of the overall

speed, by simply observing the elapsed time for the scan operation.

To mitigate the problem described above, it becomes necessary to design a benchmark suite of documents that is representative of the particular work environment. Such a suite was designed by the authors, and all major systems were evaluated using this suite of typical applications. The results of this evaluation exercise are presented later in this paper.

4.4 Document Complexity

Based on the facts discussed in the preceding subsections, it is desirable to classify documents into categories, based on factors such as complexity (text versus images, multiple fonts, etc.) and quality of printing. Such a categorization makes is feasible to think of speed and error rates in terms of the quality of the document with a specific quality.

We used a set of five classes to classify documents as shown in Table 1(a). This set, in increasing order of complexity, consists of the following classes:

<u>Class 1</u>: Text only, single column, monospaced, single pitch

<u>Class 2</u>: Text only, single column, multifont, mixed spacing

<u>Class 3</u>: Mainly text, single column, some images, any formatting of text

<u>Class 4</u>: Multicolumn document, tables

(1a) <u>DOCUMENT COMPLEXITY</u>

<u>Low Noise</u>: Original typewritten or typeset document, clearly separated characters, no skewing

<u>Medium Noise</u>: Easily readable photocopy or original laser print, characters not touching

<u>High Noise</u>: Broken and touching characters, fading ink, skewed text

(1b) <u>DOCUMENT QUALITY</u>

TABLE 1: A FRAMEWORK FOR CLASSIFYING DOCUMENTS BASED ON THEIR COMPLEXITY AND QUALITY

Class 1: Basic Text - Only Documents: All material is in a single font, with a single pitch, and with a uniform spacing. An example of this class is a typewritten document.

Class 2: Single Column Documents with Multifont and Mixed Spacing: This covers text only documents with proportional spacing and typeset and laser printed documents with multiple formats such as bold or hyphenated.

Class 3: Single Column Documents with Segregated Text and Images: Such documents contain all material in a single column format. The text is justified or hyphenated and there are some images. These images can be easily separated from the text (separate zones for text and images).

Class 4: Multicolumn Documents: Such documents contain two or more columns on a page. Apart from mostly text, there are some images and tabular material. A printed page from a newspaper will fall under this category.

Class 5: Integrated Documents: Such documents contain both text and images. A typical document of this class

contains multiple columns, and several charts or illustrations within each column.

The set of representative scanners considered in this paper was selected to include only those systems that were able to scan both images and text. However, the reading speed noted against each system is for a single column, typewritten document with uniform spacing between adjacent characters, and a single, recognizable font.

4.5 Document Quality

Since the performance of a scanner severely impacted by the quality of the input documents, it became necessary to carefully control the variations in the quality of the documents used in the benchmark tests. Document quality can be grouped into three classes as follows:

Low noise documents: This category comprises of original, typewritten and typeset documents, with normal leading and clearly separated characters. There is no (or negligible) skewing, no hyphenation, and no kerning in these documents.

Medium noise documents: This category comprises of original laser printed documents or high quality dot

matrix printed documents as well as good photocopies of such documents. The contrast is good and skewing is low (under 2%). Further, the characters do not touch each other. There may, however, be some instances of kerning, hyphenation, and uneven leading.

High noise documents: This category comprises of second or later generation photocopies with broken segments of text and characters touching each other. Usually, there is low contrast and skewed text.

The above concept of quality of documents, summarized in Table 1(b) is used to evaluate the performance of different products currently available in the market.

4.6 Recognition Techniques

Recognition technique is a qualitative variable that tries to capture the sophistication of the technique used in the recognition process. Various recognition techniques such as matrix-matching and feature extraction offer different capabilities for reading. The implication of using different recognition techniques is examined in detail in Section 5.

4.7 Man and Machine Interface

In order to minimize the total cost of scanning and

editing documents, one important factor is consider is the interface to the reading machine. A higher-speed system that requires special skills and training of dedicated operators may, at times, be less desirable that a lower-speed system with a very user friendly interface. Two of the evaluation criteria that represent this variable are trainability and docment handling. In addition, it is also important to examine the interface between the reader and other computational equipment.

4.8 Performance of the Systems

The capabilities of the various reading systems for documents of different complexity and quality are summarized in Table 2. This table compares the key specifications and makes predictions about new features.

At the high end, scanners such as Palantir CDP, Kurzweil 4000 and Kurzweil Discover 7320 were are to handle multiple columns documents as shown in Figure 1. Class 5 documents, with no clear separation between text and graphics, are handled most effectively by the Palantir CDP system. In the case of such complex documents, the process for editing and reconstitution of the original format is time consuming, exceeding ten

EVALUATION CRITERIA	CompuScan PCS 240	Data copy 730
DOCUMENT PROCESSING Scanning Speed[1] - in characters/second - in pages/minute	 110 characters per sec 15 to 20 secs.	 20 to 25 secs
Formatting (alignment, tabs)		
Multiple Column Handling	Multiple scans required	Multiple scans required
Handling of Graphics/Text on Same Document (Single pass, User-definable zones)	User specified zones	
Capability to Overlap User-Defined Zones (text <u>and</u> graphics)	No	No
Non-scan colors	Red and light pastels	
ERROR HANDLING Accuracy of Recognition	99 % for type-written matter	
Error Flagging Capability	Inserts non-recognisable character to flag errors	On-line correction and training
GRAPHICS HANDLING Resolution (in d.p.i.)	Upto 300 dpi	Upto 300 dpi
Extent of Compression	30 to 75 %	
Raster to Vector Conversion Capability for Line Drawings	Not available	Not available
TECHNOLOGY Recognition Technology	Matrix matching	Matrix matching
Actual Documents Scanned		
PHYSICAL DOCUMENT HANDLING Data Form (Text/Graphics)	Text / graphics	Text / graphics
Input Document Procedure (Flatbed or Auto)	Automatic	Flat bed
Maximum Hopper Capacity	50 sheets	1
Size Handled	8.5 X 11 in	8.5 X 14 in
RECOGNITION OF TEXT Fonts or Styles Read	12 standard fonts read can not read type-set matter	Can not read type-set matter
Quality of Documents (Photocopies, original, dot-matrix)	Can be processed	Can be processed
Point size Limitation		
Multiple Fonts in Document	Can be processed	Can be processed
Font Selection (User specified or automatic)	Automatic	Automatic
Character spacing (Mono/Proportional)	Mono spacing	Mono spacing
Character & Layout Attributes (Bold, italics, Underlining)	Can process underlined text only	Can process underlined text only
Contrast Flexibility	None	None
Trainability	Not trainable	User trainable for new fonts

EVALUATION CRITERIA	Kurzweil K4000	Kurzweil Discover 7332?
DOCUMENT PROCESSING Scanning Speed[1] - in characters/second - in pages/minute	After trainibility 35-40 characters per second	30-60 characters second
Formatting (alignment, tabs)		
Multiple Column Handling	in several passes with operator intervention	Operator definable zones using a light pen
Handling of Graphics/Text on Same Document (Single pass, User-definable zones)	one pass with user defined zones	several passes with operator specification of zones at each pass
Capability to Overlap User-Defined Zones (text <u>and</u> graphics)	no	no
Non-scan colors	all colored text is omitted	all except light pastels
ERROR HANDLING Accuracy of Recognition	good	depends on the extent of training
Error Flagging Capability	Yes	3 levels of flagging
GRAPHICS HANDLING Resolution (in d.p.i.)	Software selectable up to the reolution of the camera(346 x 667)	Up to 300 dpi
Extent of Compression	CCITT group 3	depends on the PC software formatted for
Raster to Vector Conversion Capability for Line Drawings	no	no
TECHNOLOGY Recognition Technology	"a sophisticated matrix matching" where characters and words are compared with those stored in a lexicon	Same
Actual Documents Scanned		
PHYSICAL DOCUMENT HANDLING Data Form (Text/Graphics)	Both	Both
Input Document Procedure (Flatbed or Auto)	Page feeder, automatic Feeder	
Maximum Hopper Capacity	50	10
Size Handled	Max 11" x 14"	Max 8.5" x 14"
RECOGNITION OF TEXT Fonts or Styles Read	Omnifont reader can learn to read any special or foreign character	Virtually any font
Quality of Document (Photocopy, Original, Dot matrix)	Can handle low quality documents	Quality dot matrix and good photocopies
Point size Limitation	5-24	8-24
Multiple Fonts in Document	yes	yes
Font Selection (User specified or automatic)	Not applicable	Not applicable
Character spacing (Mono/Proportional)	Any line spacing	Any line spacing
Character & Layout Attributes (Bold, italics, Underlining)	Read and restitutes layout attributes	Distinguishes onley underlined and non underlined text
Contrast Flexibility	Not applicable	16 levels of contradst
Trainability	all fonts needs training by an expert to create a customized training set	No operator intervention. AI software

[1] Scanning Speed refers ... speed taken to scan a high-quality document in a recognizable font

<u>TABLE 2</u>: CHARACTERISTICS AND CAPABILITES OF REPRESENTATIVE SET OF CURRENT PRODUCTS

EVALUATION CRITERIA	Microtek 300 C	Abaton Scan 300/SF
DOCUMENT PROCESSING Scanning Speed[1] in characters/second in pages/minute	dependent on the host IBM AT 80ch/sec IBM PC 30-35 ch/sec	
Formatting (alignment,tabs)	stores the formatting	
Multiple Column Handling	no	
Handling of Graphics/Text on Same Document (Single pass, user-defined zones)	no	
Capability to Overlap User-Defined Zones (text and graphics)	no	
Non scan colors	no	
ERROR HANDLING Accuracy of Recognition	Good for typewritten material Low for typeset material	
Error Flagging Capability	yes	
GRAPHICS HANDLING Resolution (in d.p.i.)	300	
Extent of Compression	depends on formatting	
Raster to vector Conversion Capability for Line Drawings	no	
TECHNOLOGY Recognition Technology	Feature extraction (structural analysis)	
Actual Documents Scanned		Hardware unable to scan sample documents
PHYSICAL DOCUMENT HANDLING Data Form (Text/Graphics)	Both	
Input Document Procedure (Flatbed or Auto)	Automatic feeder	
Maximum Hopper Capacity	50 pages with optional sheet feeder	
Maximum Size Handled	8.5 x 14	
RECOGNITION OF TEXT Fonts or Styles Read (No.)	Most usual fonts but limited number of fonts	
Point Size Limitation	6-12	
Multiple Fonts in Document	yes	
Font Selection (user specified or automatic)	automatic	
Character spacing (Mono/Proportional)	Typewritten material and some capabilities capabilities for proportional spacing	
Character & Layout Attributes (Bold, Italics, Underlining)	Read but do not restitute the attributes	
Contrast Flexibility	3 levels for text 8 levels for graphics	
Trainability	non trainable	

[1] Scanning Speed refers to the speed taken to scan a high quality document in a recognizable font

EVALUATION CRITERIA	Palantir CDP	IOC Reader
DOCUMENT PROCESSING Scanning Speed[1] in characters/second in pages/minute	 100 characters/second 3-4 pages/minute	 100 characters/second 3 pages/minute
Formatting (alignment,tabs)	Maintains formatting	Maintains formatting
Multiple Column Handling	Yes	Yes, but with tab separated lines
Handling of Graphics/Text on Same Document (Single pass, User-definable zones)	Single pass handles both graphics & text. 258 user-defined zones possible	Needs two passes. Vertical length of graphics/text area can be specified
Capability to Overlap User-Defined Zones (text and graphics)	Yes	No
Non-scan colors	Light pastels	Light green, yellow
ERROR HANDLING Accuracy of Recognition	Excellent	Very good in specific fonts
Error Flagging Capability	User-definable looser or stricter error flagging	None
GRAPHICS HANDLING Resolution (in d.p.i.)	75 to 400 dpi	200-400 dpi
Extent of Compression	CCITT Group 3 or 4 standard	CCITT standard
Raster to Vector Conversion Capability for Line Drawings	No	No
TECHNOLOGY Recognition Technology	AI-based proprietary technology, feature extraction with contextual analysis	Matrix-matching
Actual Documents Scanned	Documents No. 1,2,3,4, &5	Document No. 1
PHYSICAL DOCUMENT HANDLING Data Form (Text/Graphics)	Text & Graphics	Text & Graphics
Input Document Procedure (Flatbed or Auto)	Automatic Sheet Feeder	Automatic Sheet Feeder
Maximum Hopper Capacity	50 Sheets	30 Sheets
Size Handled	Max 8.5" x 14" Min 3" x 5"	Max 8.5" x 14" Min 5" x 5"
RECOGNITION OF TEXT Fonts or Styles Read Quality of document (Photocopy, original, dot matrix)	Virtually all fonts read (Hand print not read) Can read dot matrix & 2nd or 3rd photocopy	10 fonts read including Pica, Prestige & Courier Can read dot matrix but has trouble with photocopies
Point size Limitation	6 to 28 points	8 to 12 points
Multiple Fonts in Document	Yes	Yes
Font Selection (user specified or automatic)	Automatic	User specification possible. Automatic takes longer
Character spacing (Mono/Proportional)	Mono & Proportional	Mono only
Character & Layout Attributes (Bold, Italics, Underlining)	Bold, Italics & underlining read & coded	Yes, but only underlining coded
Contrast Flexibility	Operator [illegible] (3 Levels)	
Trainability	Self-training intelligent system	Not trainable

TABLE 2: CHARACTERISTICS AND CAPABILITIES OF REPRESENTATIVE SET OF CURRENT PRODUCTS (con't)

[illegible]ATION CRITERIA	CompuScan PCS 240	Data copy 730
OUTPUT HANDLING Output Form	ASCII file for text and bit mapped image files	ASCII file for text and bit mapped image files
Output Interface (Ethernet, RS232C, Hosts)	RS 232 C, IBM PC compatible hosts	RS 232 C, IBM PC compatible hosts.
Output Format (Wordprocessor, database, worksheet)	ASCII text file compatible with popular word processors	ASCII text file compatible with popular word processors
Output Medium (Standalone or host-dependent)	Host dependent	Host dependent
QUOTED PRICE Including Software	$ 6000	
REMARKS Major Advantages	Special provision for security of data	Trainable for new fonts
Major Disadvantages	Requires a lot of operator intervention	can process one sheet of paper at a time
DEVELOPMENTS EXPECTED -Near term, < 1 year		New version to process type-set matter under development
-Longer term, 1-5 years		May process all type of fonts
Other Comments		
LOCAL CONTACT	D C Victor Yuan SYNTECH INC	D C Victor Yuan SYNTECH INC

[illegible]ATION CRITERIA	Kurzweil K4000	Kurzweil Discover 73320
OUTPUT HANDLING Output Form	ASCII text & bit-mapped images	ASCII text & bit mapped images
Output Interface (Ethernet, RS232C, Hosts)	Virtually any computer	IBM PC AT or compatibles
Output Format (Wordprocessor, database, worksheet)	ASCII FILE and customized format	ASCII file and common PC wordprocessors
Output Medium (Standalone or host-dependent)	Host dependant	Host dependant
QUOTED PRICE Including Software	25 000$	10 000$
REMARKS Major Advantages	-Omnifont reader -Good ability to read low quality documents, and foreign and special chracters - Reasonable ability for handling of technical documentation Good ability of adjustment for skewness	-Omnifont reader Good ability to read several kind of documents - User friendly interface - User definible lexicon of 10000 words
Major Disadvantages	- Training of the system is time consuming Requires a dedicated opertaor Limited ability for graphics - The size of the system makes it non-sharable - Low scanning speed	Limited ability for handling of multiple columns and mixed graphics and text Scanning speed is limited
DEVELOPMENTS EXPECTED Near term, < 1 year	No or very limited improvement	Some improvement for the software
Longer term 1-5 years	Limited improvement	Reduction in the price and important improvement in the the accuracy
Other Comments	The product has been in thenas been in the market for several years and theis based on research done in the mid-seventies. Few changes in the product occured since its entry in the market	The hardware part of the product is the same as for the Abaton and Microtek scanner
LOCAL CONTACT	Swain Mills Kurzweil Research Center 185 Albany St Cambridge, MA 02139	

<u>TABLE 2</u>: CHARACTERISTICS AND CAPABILITIES OF REPRESENTATIVE SET OF CURRENT PRODUCTS (con't)

EVALUATION CRITERIA	Microtek 300 C	Abaton Scan 300/SF
OUTPUT HANDLING Editing Program	editing in common PC wordprocessors	
Output interface (Ethernet, RS232C, Hosts)	IBM and compatibles, Macintosh	
Output Format (Wordprocessor, database, worksheet)	Common wordprocessors	
Output Medium (Standalone or host-dependent)	host dependant	
QUOTED PRICE including Software	2500$	
REMARKS Major Advantages	very good graphic capabilities Good performance of the software considering the price good user interface	
Major Disadvantages	Limited capabilites for typeset and proportionally spaced material Poor ability of multiple column and mixed text and graphics documents	
DEVELOPMENTS EXPECTED Near term < 1 year	Improved reading of typeset material	
Longer term 1-5 years	Improved accuracy - Could become an excellent tool for desktop publishing antd the low to medium rANGE MARKET	
Other Comments		
LOCAL CONTACT	Ken Barber Provue Inc 145 South St Boston MA	Ann[illegible] Sherman Howe Computer Center [illegible] Cedar Street Wellesley [illegible] MA [illegible]

EVALUATION CRITERIA	Palantir CDP	IOC Reader
OUTPUT HANDLING Output form	ASCII text & bit-mapped images	ASCII text & bit mapped images
Output interface (Ethernet, RS232C, Hosts)	Ethernet, IBM PC, & Sun	IBM PC/XT AT & other MS DOS compatibles
Output Format (Wordprocessor, database, worksheet)	Wordprocessor, database or worksheet & graphics editing	Wordprocessor & graphics editing
Output Medium (Standalone or host-dependent)	Host-dependent	Host dependent
QUOTED PRICE including Software	$ 48 000 including software & Ethernet interface	$3 500
REMARKS Major Advantages	i Exceptional print reading capability ii Capability to process tables & forms iii Suitable for intensive production use	i Specific font reading capabilities can be developed by the company to order ii Excellent for half tones 128 shades of grey available
Major Disadvantages	i Cannot auto-specify graphics/text areas on document ii Limited half toned graphics capability	Cannot read typeset material
DEVELOPMENTS EXPECTED Near term < 1 year	i Networking to other scanners & multiple hosts ii Drop in price (?)	
Longer term 1-5 years	i Raster to vector conversion for line drawings (?) ii Advanced AI techniques for natural language parsing leading to error free recognition (?) iii Automatic identification of text/graphics areas on document (?)	Typeset reading capability
Other Comments		
LOCAL CONTACT	Robert Warner Palantir Corporation 3 New England Executive Park Burlington MA 01803	Bob Haskell System Automation Inc 15 Lakeside Office Park Wakefield, MA 01880

TABLE 2: CHARACTERISTICS AND CAPABILITIES OF REPRESENTATIVE SET OF CURRENT PRODUCTS (con't)

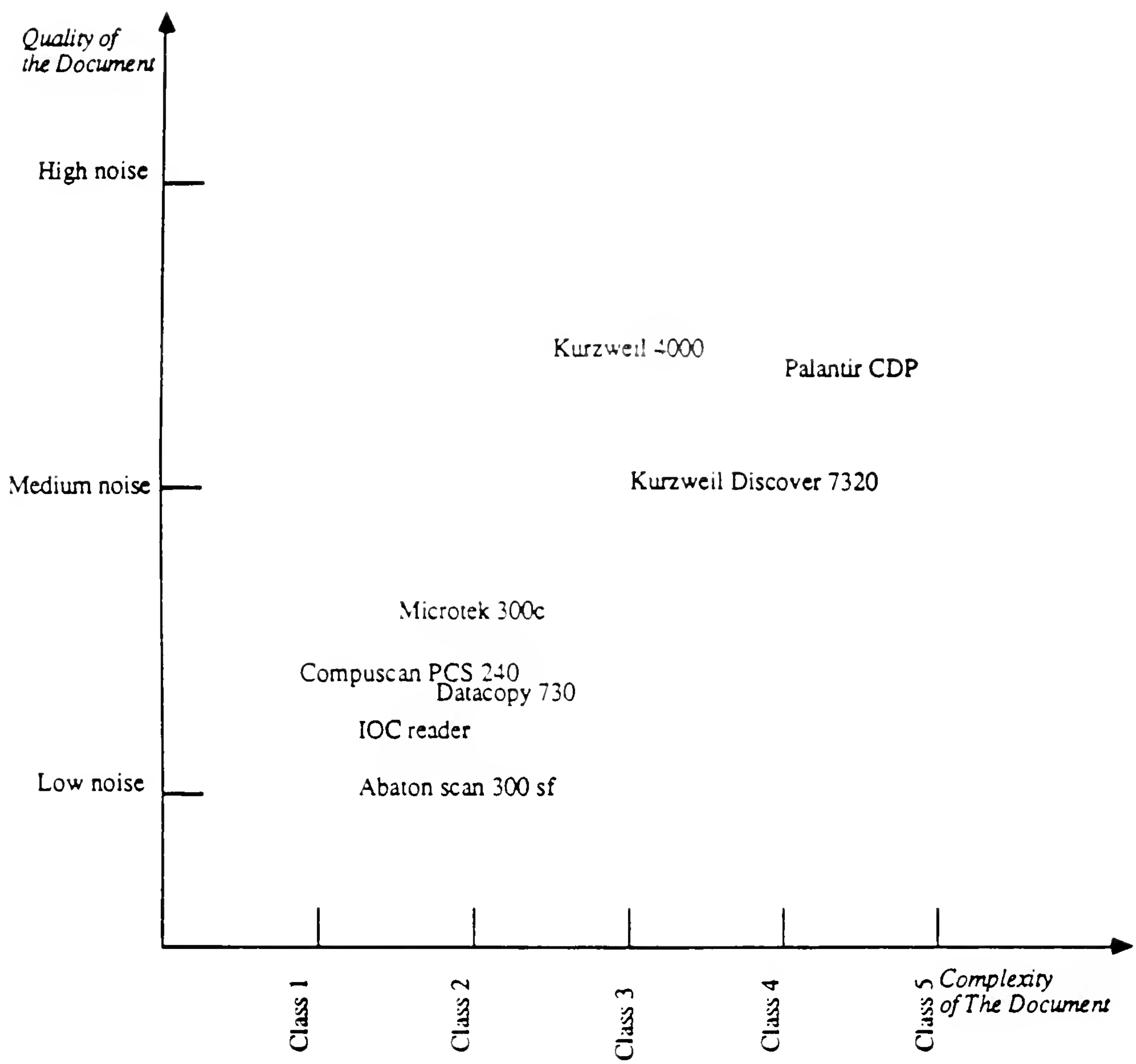


FIGURE 1: CAPABILITIES OF CURRENT PRODUCTS ON DOCUMENT COMPLEXITY/QUALITY MATRIX

minutes for the sample used in the benchmark study.

This incidence of errors in the high-end scanners related to the quality of the document. Characters that touch each other and with broken strokes constitute the major sources of errors. Even a typewritten document caused a significant number of errors in all the systems that were studied in cases where the quality of the document was low. A single handwritten mark or even a speck of dirt is a potential source of error for the reading mechanism employed in most scanning systems.

Many of the problems discussed in this section can be mitigated using the techniques described in the next section.

5. RECOGNITION TECHNIQUES

The recognition of text, the scanning of images, and the raster to vector conversion of technical drawings have usually been considered independently of each other. The technologies corresponding to these three areas must be integrated together to accurately scan and process complex technical documentation. One framework that allows recognition of both text and images is presented in this section.

The three major stages in the processing of a document are preprocessing, recognition, and post-processing. These are discussed in the following subsections.

5.1 Preprocessing

Preprocessing is the conversion of the optical image of characters, pictures, and graphs of the document into an analog or digital form that can be analyzed by the recognition unit. This preparation of the document for analysis consists of two parts: image analysis and filtering.

5.1.1 Image Analysis

The first stage of image analysis is scanning. Scanning provides a raster image of the document with sufficient spatial resolution and grey scale level for subsequent processing. In the case of a picture or a graph, the latter issue of grey scale level is more important that in the case of text. For text, this phase consists of locating character images. With the exception of high end scanners such as the Palantir CDP and to some extent the Kurzweil Discovery 7320, which employ contextual analysis as well, reading machines are character-oriented. Each character is treated as a unique event and it is recognized independent of

other characters. This implies that the document must be first segmented into separate characters, and then the identity of each character recognized.

The optical system first takes a raster image of the area that is supposed to enclose the character. Alternatively, the raster image representing the character is cut out of the image of the document. In either case, the image is transmitted sequentially to a single character recognition subsystem. If the image, or the information on the features of the character constituting the image, possesses characteristics which are significantly different from the characteristics maintained by the character recognition subsystem, then the particular area is deemed to be either an unrecognized character or noise. Depending on the system, the output is expressed as a flag or a blank. In some systems such as those manufactured by Compuscan and Microtek, any graphic or character input which is outside the size limitations is not flagged but skipped. The formatted output, in such a case, contains a blank zone corresponding to the input of the improper size.

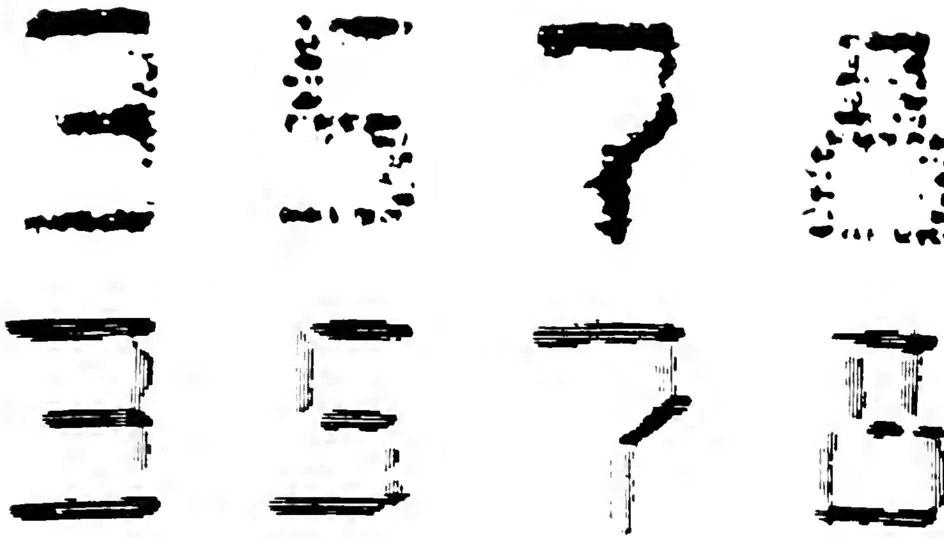
5.1.2 Filtering

Filtering minimizes the level of noise. The latter may be caused either in the source document or by the opto-electrical transformation mechanism. The process of filtering also enhances the image for easier recognition. One filtering process that eases the extraction of the features of the character in the recognition phase has been recently proposed independently by Lashas (3) and Baird (4). They present two OCR readers in which black marks constituting the character are transformed into quantized strokes. Their approach is depicted in Figure 2.

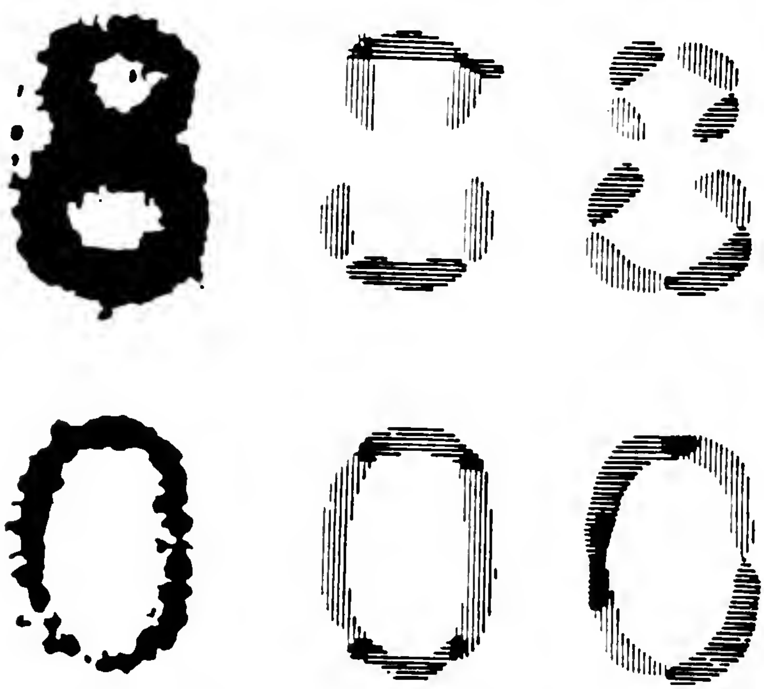
5.1.3 Preprocessing: New Approaches

The preprocessing phase consists of deriving a high level representation of the contents of the image. The scanned document is seen as a set of blocks corresponding to independent ways of representing information such as text, line drawings, graphs, tables, and photographs. Understanding of this document involves the following:

- Identification of the major blocks of information.



HORIZONTAL AND VERTICAL QUANTIZED STROKE SEGMENTS



FOUR DETECTED DIRECTIONS OF QUANTIZED STROKES

FIGURE 2: THE APPROACH OF QUANTIZED STROKES

- Identification of the spatial relationships between the different blocks (for example, the reading order so that the logical connections between different blocks of text or graphics can be easily derived).
- Identification of the layout features (number of columns and margins, and justification.
- In the case of text, further identification of headlines, footnotes, etc.

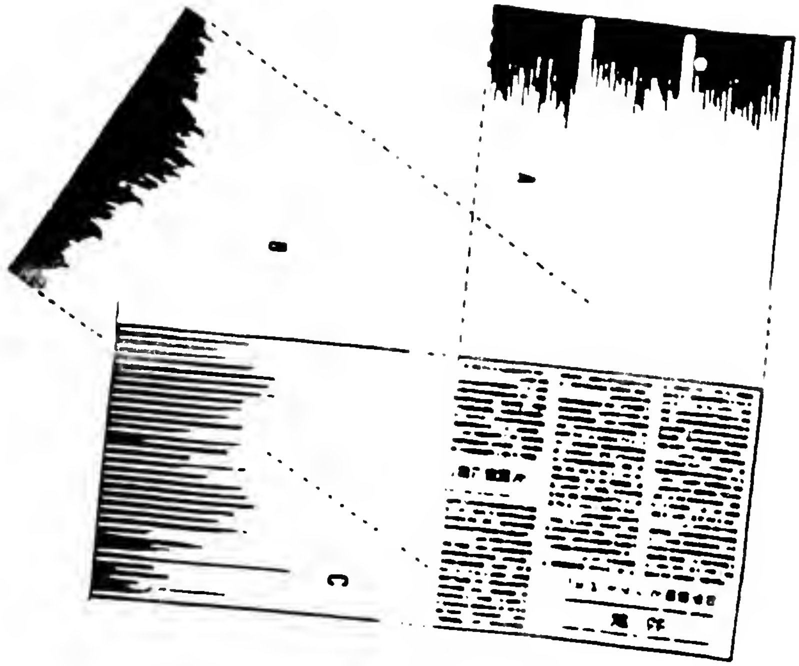
Typically, text is first distinguished from other information, and then columns of text are recognized. These columns are next split into lines of text which, in turn, are segmented into single character images. A reader which is able to accept a free format document was described by Masuda et al (5). Their scheme of area-segmentation uses projection profiles obtained by projecting a document image on specific axes. Each profile shows the structure of the document image from a particular angle. The projection profile is very sensitive to the direction of projection. In order to check for skew normalization, the image of the document is incrementally rotated and the horizontal and vertical projection values are noted. Through an

analysis of the intensity of these values and an examination of different areas, general text and headlines are separated (Fig. 3). Another reading system based on this approach is described in (6).

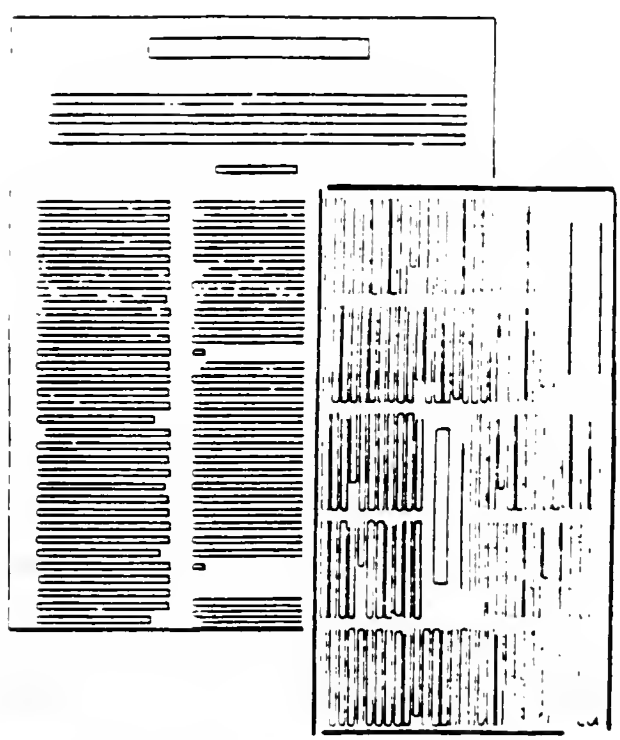
Currently available commercial system allow only manual area segmentation. While such segmentation methods are fairly rudimentary in the low-end scanners, methods that permit the operator to define several text and graphics "windows" within a page are available on products from Palantir and Kurzweil. The Palantir CDP allows the use of a "mouse" to define up to 256 text, graphic, or numerical zones. The Kurzweil 4000 system enables the identification of graphic and text zones through the use of a light pen by the operator.

5.2 Recognition

Recognition occurs at the level of characters in most commercial page readers. However, high end scanners such as the Palantir CDP are now complementing character recognition by sophisticated contextual analysis. Techniques used for character recognition and contextual analysis are discussed in the following paragraphs.



(a) PROJECTION PROFILES



(b) TEXT LINE AREAS EXTRACTION

FIGURE 3: A STRATEGY OR RECOGNIZING INFORMATION FROM CLASS 5 DOCUMENTS

5.2.1 Character Recognition

The character recognition problem is essentially one of defining and encoding a sequence of primitives that can represent a character as accurately as possible. The most common approaches to character recognition are described below.

5.2.1.1 Template matching technique

Among the oldest techniques for character recognition, template matching involves comparing the bitmap that constitutes the image of the character with a stored template. The amounts of overlap between the unknown shape and the various stored templates are computed and the input with the highest degree of overlap is assigned the label of the template.

The primitive in this case is a very simple function that assigns one value to a point of the bitmap if it is black and another if it is white. The performance of different readers using

this technique depends on the decision algorithms. This method offers high-speed processing especially for monofont pages. For example, the document reader made by Scan-Optics can process up to 2000 characters per second. Even in the case of documents containing several fonts which are extracted from a limited set of fonts of a given size, this techniques can be very effective. Moreover, this method is relatively immune to the noise generated during the opto-electrical transformation.

However, the method is less effective in the case of unknown fonts and multifonts and multisize characters. Further, this method imposes constraints on the format of the page in areas such as the spacing of the letters and the position of the text. If these constraints are relaxed, template matching becomes slow and costly, because of the need to translate, scale, and rotate input images.

5.2.1.2 Feature extraction

In contrast to the template matching technique which emphasizes the importance of the overall shape, the feature extraction approach focuses on the detection of specific components of the character. This approach assumes that local properties and partial features are sufficient to define every character.

Feature extraction techniques identify local aspects, such as pronounced angles, junctions and crossing, and define properties such as slope and inflection points. In one method of recognition, a boolean or a numerical function that characterizes each feature is calculated and then applied to the given image. Another method involves definition of a partial mask that can be displaced systematically positioned on the pattern. In a more recent strategy proposed in (5), recognition is based on the analysis of the direction and connective relationships of the strokes

of the character.

5.2.1.3 Structural analysis methods

The use of structural analysis methods is a recurrent theme in literature (2, 3, 4, 5, 7). In this method, each character is defined as a set of topological primitives such as strokes, segments, holes, and arcs.

Isolated, unbroken characters are first ecognized using a structural description. These descriptions are independent of the position and the size of the character. A parametrization of the shape is then performed so that the results of the structural analysis can be compared with a stored list of shapes. Next, the shapes are clustered to discriminate between characters and to classify them. The power of structural analysis depends of the number of features of each character used by the system. An example of the structural analysis of a character is shown in Fig. 4.

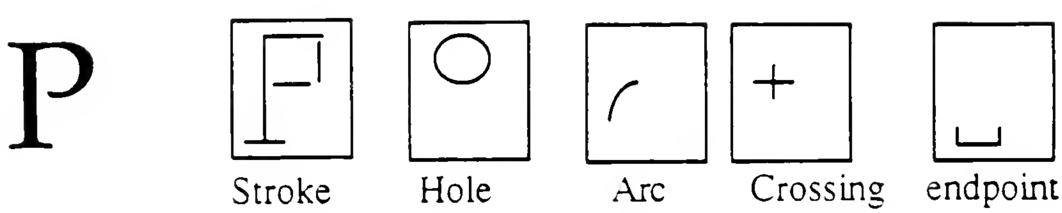


FIGURE 4: EXAMPLE OF STRUCTURAL ANALYSIS OF A "P"

Methods of structural analysis are frequently utilized in commercial reading machines. The products made by Microtek, Palantir and Kurzweil make use of structural analysis methods in their recognition process.

5.2.1.4 Character Classification

So far, the discussion has centered on single characters in a single font. Most documents contain multiple characters with considerable variation across fonts. A statistical approach is used for dealing with these variations. In multifont reading machines, tests of character recognition are based on statistical data constituted by a training and test set. An omnifont reader like the Palantir CDP implies that it has been trained on a large set of fonts.

Feature extraction is generally complemented by classification. In order to be classified, characters have to be discriminated from each other. The

classifier is designed to construct a discriminant function underlying a complex probability distribution taking into account the non-linear variations in the characters.

The two essential steps in recognition, feature extraction and classification, have different optimization criteria. Shurman [2], feels that feature extraction should be optimized with respect to the reconstruction of the character, while classification should be optimized with respect to recognition. Consequently, feature extraction should not be directed toward recognition through definition of features that minimize the classification process.

Classifiers are usually adapted to different fonts by the manufacturer. However, some machines like the Kurzweil 4000 or Datacopy 730 provide a classifier which allows for the adaptation of new fonts by the operator.

An on-line training capability for multifont recognition can also be provided [2]. This concept of trainability is similar to that employed in modern speech recognition systems which overcome the variability of voice between speakers by requiring that adaptation be performed during a training phase prior to regular operation. The main disadvantage of trainable systems lies in their slow training speed as the training set contains several thousands of characters.

5.2.1.5 Limitations of Current Recognition Techniques

Ideally, feature extraction techniques should be able to generate a set of characters that are independent of font and size. However, the wide variety of fonts encountered in office environments results in a huge number of possible characteristics. Furthermore, ambiguities occur due to the similarity in the features of different characters

across fonts. The distinction, for example, between 1 (one) , I (capital I) and l (el) across and even within fonts is not obvious. Moreover, there are a number of characters which cannot be correctly identified without information about their size or position. For example, the character "O" (capital o) in one size corresponds to the lower case "o" in a larger size; further, this character is the same as 0 (zero) in several fonts.

These factors seem to imply that recognition techniques, despite involving heavy computation, are prone to recognition ambiguities. Solely relying on the physical features of the characters results in a high rate of errors and rejects. To combat this problem, recognition methods are complemented by contextual and syntactical analysis, aided by customizable lexicons, especially in high-end systems.

In the case of low quality documents with different kinds of noise and broken and touching characters, the error rate is high. Contextual analysis becomes a necessary condition for reducing the rate of errors and rejects in such situations.

5.2.2 Separation and Merged Characters

The breaking of images into character blocks is based on the assumption that each character is separated by horizontal or sloped line of blank spaces. However, in the case of tight kerning, inadequate resolution of the scanner, poor quality of the document, or high brightness threshold, adjacent characters spread into each other.

Separating merged characters involves three processes, as follows:

- Discriminating multiple characters from single characters blobs;
- Breaking the blobs into components that are identified as single character blobs;
- Classifying the different characters and deciding whether to accept or reject the

classification.

Defining a set of criteria for distinguishing between adjacent characters is difficult because of the many ways in which characters merge together and the fact that merged characters contain misleading strokes. Separation of characters is a computationally intensive task that significantly slows the overall recognition process. Only the Palantir CDP and the Proscan systems offer some abilities for automatic separation of merged characters. Separation is possible in a limited number of cases and for pairs of characters only.

5.2.3 Contextual Analysis Techniques

Contextual analysis is of two types; layout context analysis and linguistic content analysis. Layout context analysis covers baseline information on the location of one character with respect to its neighbors. It generates, for example, formatting information and is usually language-independent. Linguistic analysis, on the other hand, includes spelling, grammar and punctuation rules. For example, a capital letter in the middle of a word (with lower case

neighbors) is not accepted. Layout context analysis capabilities are available on several systems available today. However, only the high-end systems offer some degree of linguistic analysis capabilities.

5.2.3.1 Dictionary Lookup Methods

Several contextual methods for word recognition are based on the analysis or the comparison of a word or a string of characters with stored data for type and error correction. Spelling checkers are commonly used to complement character recognition devices. Such checkers may be part of the recognition software as in the case of the Palantir CDP and the Kurzweil 4000 or they may be part of the text composition software that the operator uses for correcting the processed document. In either scenario, when the size of the dictionary is large, the search time can be very long. Furthermore, if the contextual information is not considered, the scanned word may be incorrectly converted into another. Several high

speed correction methods are based on the use of similarity measures based on the fact that most errors are due to character substitution, insertion, or deletion. The similarity between two words (the correct word and the garbled word) of equal length is measured using the Hamming distance. The Modified Levenstein distance [9] generalizes the similarity measure for substitution, insertion and deletion. Minimization of this distance forms the optimization criteria. Tanaka [8] proposed two methods that yield 10 - 35% and 35 - 40% higher correction rates than a typical dictionary method and reduce computing time by factor of 45 and 50 respectively. One of these methods is based on the assumption that different characters can be classified into classes or groups that are independent of each other so that a character in one class is never misrecognized as a character in another class. This categorization helps to significantly reduce the probability of errors.

5.2.3.2 Use of Statistical Information

The frequency of occurrence of different characters is different. For example, the letter "e" has the highest frequency of appearance in an English document. Further, several character combinations and sequences of appearance are more likely to appear than others. For example, the letter "t" is the most frequent first letter of a word, and the letter "q" is always followed by a "u." The frequency of occurrence of a character within a string of given text can be efficiently modeled by a finite-state, discrete-time Markov random process. The use of statistical distributions of different combinations of N characters (N-grams) allows the implementation of error correction algorithms such as the Viterbi algorithm. This algorithm can reduce error rate by one half according to Hou [9].

5.2.3.3 Linguistic Context Analysis

Characters are primitives of strings

constrained by grammatical rules. These rules define legitimate and non-legitimate strings of characters. Based on the recognition of some words, the class (noun, verb, etc.) to which they belong can be identified along with the applicable syntactic analysis to identify misspellings, misplacements, and other syntactic errors. A string of words or a sentence can be decomposed using a parsing tree. There are several efficient parsing trees for different types of grammatical structures.

Syntactic analysis assumes that the text is constrained by one type of grammar. This assumption need not hold in all the cases. Several technical documents contain uncommon words: serial numbers, technical words, or abbreviations. Such situations require interaction between the technical operator and the system or the use of dictionary lookup methods. Except in such special situations, linguistic context analysis techniques can be utilized to identify and to

correct reading errors.

5.2.4 Future of Recognition Technology

Conventional techniques for character recognition based solely on geometric or analytical properties are not sufficient for processing complex documents with good accuracy and at high speed. While the use of contextual analysis improves the accuracy, it does not increase the processing speed. To improve speeds, it becomes necessary to analyze the document simultaneously from several points of view. For example, one analysis unit can process characters, another can perform contextual analysis of the text, and a third can analyze the image. This approach has been adopted, for example, in the Palantir CDP system which contains five Motorola 68000 microprocessors.

Most researchers deem that the use of a single recognition techniques is insufficient to solve all ambiguities in the recognition process. As such, general algorithms must be complemented with handcrafted rules for special cases, such as for distinguishing letters that can be easily confused (such as "o" and "O" and "6" and "b"), and for

recognizing characters printed in pieces (such as "i", "j", and ";"). Shurman [2] sees an ideal system as one supervised by a process control unit that collects all the pieces of information from the subsystems, orders them, and then makes appropriate final decisions.Such a strategy, which is functionally depicted in Figure 5 allows the recognition of different parts to be done in parallel.

5.3 Postprocessing

After the text and graphics are recognized, they must be encoded for transmission to, and processing by, a word processor, a graphic editor or a desktop publishing program. In the case of text, the image is most commonly converted into the ASCII format. In the case of graphics, the storage requirements for the document vary greatly, depending on the extent of compression. The storage capacity may pose a major constraint in some cases. For example, in systems which scan at 300 dots per inch (dpi), one single 8.5" x 11" page requires 900 kilobytes of memory if stored in raw bitmap format. While the advent of optical disks with storage capacities in gigabytes would tend to mitigate this problem to some extent, it is usually desirable to use more efficient storage strategies.

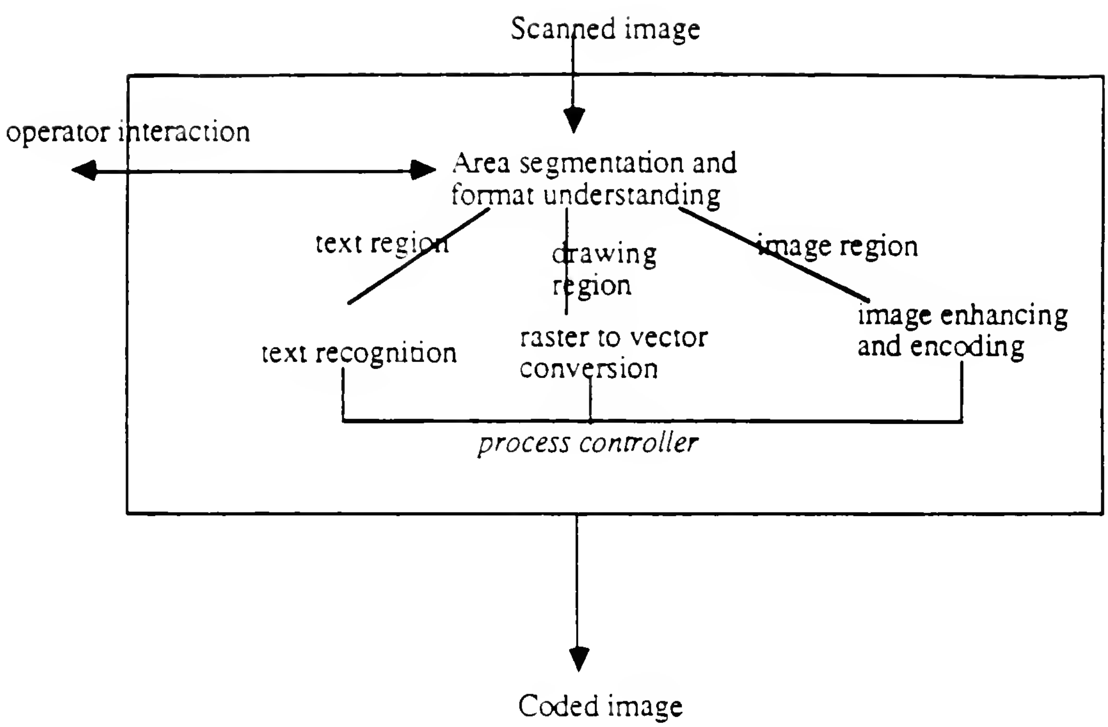


FIGURE 5: A MEDIA PROCESSING STATION

6. TRENDS AND PROJECTIONS

A number of diverse forces are driving the scanning industry. Apart from the rapid advances in hardware and software technology, that have resulted in the cost survey shown in Figure 6, there are demands from a number of sectors. Defense, financial, engineering, legal and medical sectors represent some of the potential major users of scanning technologies. Another significant force is from the area of desktop publishing. The impact of this industry and other factors is discussed below.

6.1 Influence of Desktop Publishing

Many vendors view the desktop publishing industry as a good candidate for the wide application of scanners. Several companies (Abaton, Datacopy and others) now offer OCR software at a price of a few hundred dollars per copy, aimed towards individuals charged with producing in-house publications. These readers, which constitute the low end of the scanner market, are able to handle the most commonly used fonts. More versatile packages capable of reading a larger number of fonts at still lower prices are gradually appearing in the market, leading to greater sophistication and ease in the processing of complex documents for desktop publishing applications.

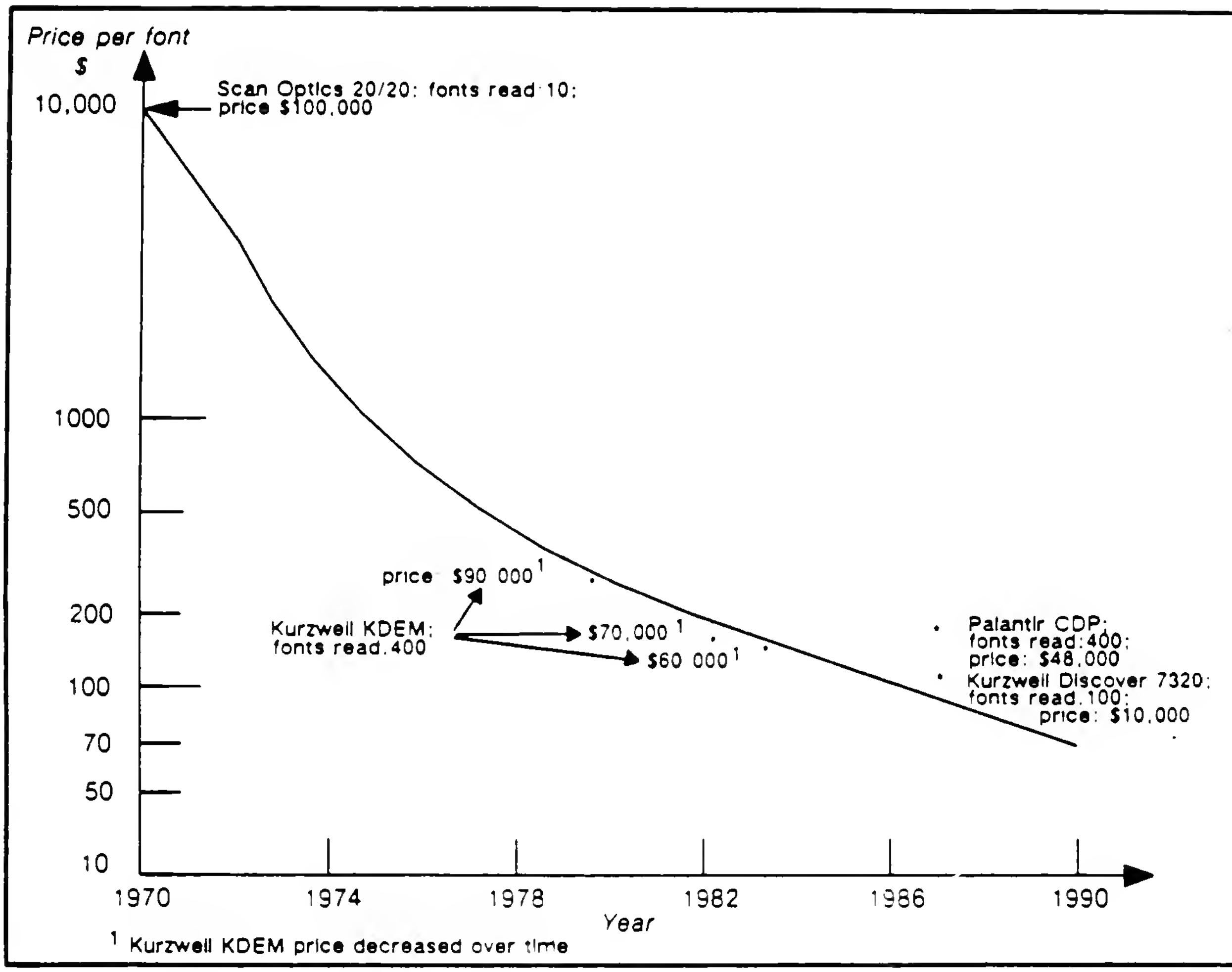


FIGURE 6: DECLINING SYSTEM COSTS PER FONT (1970 - 1990)

Users in the desktop publishing market need automatic recognition of the fonts.

Consequently, machines that require training such as the Kurzweil 4000 or the INMovatic Star Reading system are experiencing difficulty in accessing this market. Trainable systems are relevant only for specialized needs. The Discover 7320 system, launched by Kurzweil, is targeted towards these needs and represents the high end of the desktop scanner market. This is the direction in which future scanners are likely to evolve.

6.2 Coalescence of Text, Image and Graphics Processing

Graphical information can be stored either as simple bitmaps or as a set of standard geometric entities. The latter option allows for easy editing and storage in a library of symbols. Currently available scanners do not offer these capabilities. However, developments are taking place in the area of Computer Aided Design (CAD) for converting raster images of line drawings into vector graphics files, which can be easily modified with graphic line editors. Raster-to-vector conversion systems are now available from several companies including Houston Instruments and Autodesk Inc. However, at present only approximate solutions

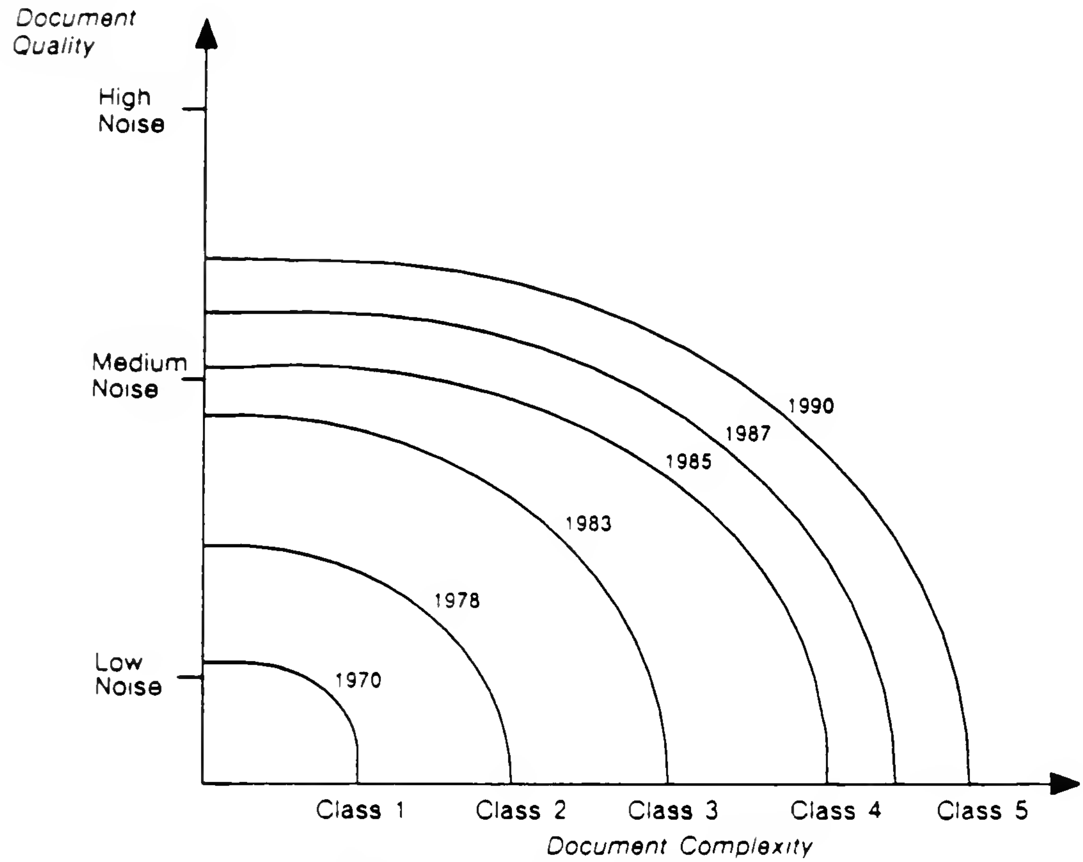
are supported. For example, all curves are decomposed into line segments. It is expected that ideas from the arena of raster to vector conversion will be combined with scanning technologies to enable text, images, and line graphics to be all edited and processed through a composite package. The overall trend is shown in Figure 7.

6.3 Integration of Facsimile and Character Recognition

Character recognition will be increasingly integrated into facsimile equipment in order to provide access to facsimile data on the computer. At the same time, facsimile capabilities are also becoming available on document readers. Datacopy, for example, allows their 730 scanner to be used as a fax machine for transmitting data.

6.4 Integration in the Automated Office

The reading machine will no longer serve as a standalone unit. Instead, its capabilities will be integrated into the office workstation. Palantir, for example, recently introduced the "Recognition Server" which constitutes the recognition unit without a scanner. This enables, for example, the combination of the recognition unit with storage devices to support automatic indexing for image-based storage and



DOCUMENT COMPLEXITY

Class 1: Single column, monospaced, single pitch
Class 2: Single column, multifont, mixed spacing
Class 3: Single column, some images, any formatted text
Class 4: Multicolumn document, tables
Class 5: Multicolumn, mixed text and images

DOCUMENT QUALITY

Low Noise: Original typewritten or typeset document, clearly separated characters, no skewing
Medium Noise: Easily readable photocopy or original laser print, characters not touching
High Noise: Broken and touching characters, fading ink, skewed text

FIGURE 7: EVOLUTION OF SCANNING CAPABILITIES

retrieval. Currently restricted by storage space limitation and retrieval delays, large scale image-based storage and retrieval will become a viable proposition when advanced storage technologies like optical discs and CD-ROMs are combined with document processing units like the Recognition Server.

6.5 Networking

The recognition system will become a shared resource on a network of work-stations, reducing the cost and increasing the document-processing throughput. Here again, Palantir's new product allows each host to be accessed and controlled through specific software. This reduces the amount of human intervention, allowing greater efficiency in the automated office.

Japan's NTT Laboratory, for example, has proposed a system consisting of several workstations, a file station and a media processing station where the recognition units are located. A document scanned at a workstation is transmitted to the media processing station which converts it into the appropriate format. All scanned documents are stored at the file station as shown in Figure 8. Errors can be rectified from any workstation at the convenience of the operator.

6.6 Impact of Artificial Intelligence Techniques

Advances in Artificial Intelligence techniques in areas related to character recognition and image analysis will lead to faster, more accurate and versatile reading machines. Semantic analysis and natural language parsing aids for context analysis will lead to better identification of letters and words, reducing both the error rates and the reject rates as well as the need for operator intervention. Reading machines will be able to read virtually all printed material.

6.7 Use of Special Purpose Hardware

A high accuracy omnifont reader requires sophisticated algorithms for vectorization and classification, separation of merged characters, contextual analysis, and syntactical analysis. These algorithms can benefit from microprogramming and special purpose hardware optimized for specific applications. The most sophisticated readers available on the market (Palantir CDP and Kurzweil Discovery 7320) use special purpose hardware. Another example is an experimental omnifont reader designed by Kahan (1987) in which special hardware has permitted implementation of optimal asymptotic time complexity.

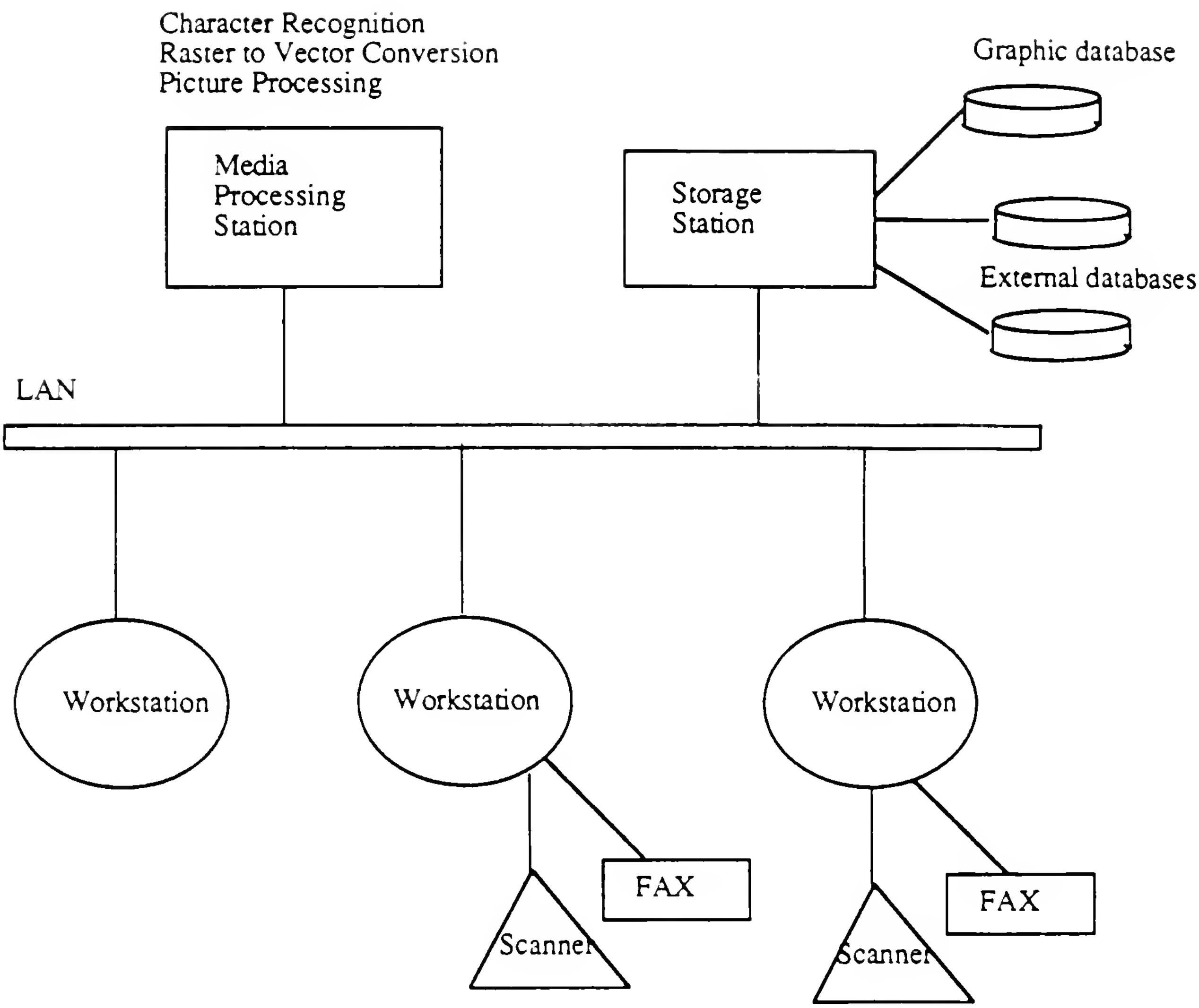


FIGURE 8: INTEGRATION OF READING MACHINES INTO INFORMATION SYSTEMS

6.8 Limiting Factors

Although it appears that reading machines will increasingly provide the ability to automatically process all kinds of documents with ease, there are several factors which are inhibiting the applicability of reading machines. These factors are as follows:

1. Although systems with higher accuracy are being designed, the accuracy provided by current systems is still inadequate for many applications. Editing of errors and the presence of unrecognized characters continue to be major bottlenecks. The time taken by the operator to overcome these bottlenecks limits the throughput of the overall system.

2. Wide acceptance of the new technology will occur only after it becomes possible to handle complex documents containing both text and graphics, without operator intervention. The requirement for a dedicated operator for editing continues to make use of reading machines into a less palatable alternative.

3. Low quality documents with broken strings and touching characters constitute a major hurdle.

Such documents cannot be handled even by the most sophisticated machines. Further, reading of hand-printed characters is at a primitive stage. Major developments in syntactical and semantic analysis are needed before reading machines realize their full potential. All the major trends and projections are summarized in Table 3.

6.9 Conclusion

The field of automatic transfer of information of paper documents to computer accessible media is witnessing a lot of attention. Many commercial products are available for reading simple documents of low complexity. The emergence of new and more powerful products is encouraging many organizations to investigate the use of the nascent technology for their respective applications. At the same time, the limited functionality of current products and techniques makesit important to exercise adequate caution while identifying applications for using the new technology.

	1975	1982	1987	PROJECTED (within 1-3 years)	PROJECTED (within 8 years)
CESSING ABILITY	Text (characters)	Text plus images (ASCII text and bit maps)	Text plus images (ASCII text and bit maps)	Text and images; Separate abilities to deal with vector graphics	Text and images plus vector graphics (vector graphics files)
ITS	Typewritten text only	Most printed as well as typewritten fonts	Virtually all printed material	All printed material	All printed fonts including some handwritten text
GE LYSIS	Matrix matching for character recognition of selected fonts	Feature analysis of letters within the text, and matrix matching techniques for identifying characters	Context analysis spelling checking (lexicons), feature extraction for identifying characters	Natural language syntax analysis based techniques for text recognition; aids for distinguishing text and graphics	Semantic analysis, natural language parsing aids for context analysis and correct interpretation of letters and automatic identification of text and image areas
TING	Simple editors for correcting correcting errors within data scanned	Text editors and word processors for editing pages	Separate editing facilities for text (text) editors/word processors) and images (through graphic editors)	Limited degree of integrated editing	Integrated editor for text, images, and vecto graphics with interface for other packages like DBMSs
LIZATION EAS	Document processing with no editing requirements	Documents with text processed through text editors	Pages scanned and processed through word processors, and graphic editors	Documents with text and graphics processed	Customization of printed page processing software
ROR TES	2-3 percent for a good quality document	2-3 percent for a typewritten page	0 1 percent for a typewritten page	0 5 percent for a typewritten page, 1 percent for a typeset page	0 percent for a typewritten page. 0 1-0.4 percent for a typese page

TABLE 3: SCANNERS - TRENDS AND PROJECTIONS

REFERENCES

[1] Ullman, J. R., "Advances in Character Recognition," in Applications of Pattern Recognition, Editor: K.S. Fu, CRC Press, Inc., Boca Raton, Florida, 1982, p. 230.

[2] Shurman, J., "Reading Machines," Proc. of International Conference on Pattern Recognition, 1982, pp. 1031 - 1044.

[3] Lashas, A. et al., "Optical Character Recognition Based on Analog Preprocessing and Automatic Feature Extraction," Computer Vision, Graphics and Image Processing, Vol. 32, 1985, pp. 191-207.

[4] Baird, H., Simon, K., and Pavlidis, T. "Components of an Omnifont Reader," International Symposium on Pattern Recognition, 1986, 344-348.

[5] Masuda I., et al., "Approach to a Smart Document Reader," Conference on Computer Vision and Pattern Recognition, 1985, pp. 550-557.

[6] Kida, H. et al., "Document Recognition System for Office Automation," Proc. of International Symposium on Pattern Recognition, 1986.

[7] Suen, C.Y., Berthod, M., and Mori, S., "Automatic Recognition of Handprinted Characters - The State of the Art," Proc. of the IEEE, Vol. 68, No. 4, pp. 469 - 487.

[8] Tanaka, Eiichi; Kohasigushi, Takahiro; Kunihiiko Shimamura. "High Speed string correction for OCR," Proc. of International Conference on Pattern Recognition, 1986, 340 - 343.

[9] Hou, H. S., Digital Document Processing, John Wiley & Sons, 1983.